Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 163-166 recibido 29-03-2017 revisado 12-05-2017 aceptado 18-05-2017

# Desarrollo de un Sistema de Segmentación y Perfilamiento Digital

## *Development of a Digital Segmentation and Profiling System*

**Jaime Vargas-Cruz[1], Alexandra Pomares-Quimbaya[1], Jorge Alvarado-Valencia[1], Jorge Quintero-Cadavid[2], Julio Palacio-Correa[2]**
[1]Pontificia Universidad Javeriana
110231, Bogotá Colombia
[2]Servicios Nutresa
050024, Medellín Colombia
[1]{jaimevargas, pomares, jorge.alvarado}@javeriana.edu.co [2]{ jquinteroc, jcpalacio}@ serviciosnutresa.com

**Resumen:** El objetivo principal de este artículo es presentar el Sistema de Segmentación y Perfilamiento Digital (SSPD), el cual, a partir del análisis de la información publicada por usuarios de redes sociales, permite perfilarlos y segmentarlos. Para lograr su propósito SSPD aplica técnicas de procesamiento de lenguaje natural, análisis de grafos y técnicas de aprendizaje automático que le permiten generar variables de tipo demográfico, psicográfico, comportamental y sociográfico para describir a los usuarios que generan publicaciones. Para garantizar el entendimiento de los perfiles y segmentos generados, SSPD proporciona un modelo de visualización interactivo que incluye una vista estática y otra dinámica en el tiempo. El sistema SSPD está siendo aplicado en Colombia usando la red social twitter; sin embargo, su arquitectura flexible permite llevarlo a otros países de habla hispana e integrarlo a otras redes sociales.
**Palabras clave:** Segmentación, perfilamiento, redes sociales, procesamiento de lenguaje natural

**Abstract:** The main objective of this article is to present the Digital Segmentation and Profiling System (SSPD), whose goal is to profile and segment users in social networks based on the analysis of information published by them. To achieve its purpose, SSPD applies natural language processing techniques, graph analysis and machine learning techniques to generate demographic, psychographic, behavioral and sociographic variables that describe the users on the network. To ensure the understanding of the profiles and segments generated, SSPD provides an interactive visualization model that includes a static view and a dynamic view over time. This system is being implemented in Colombia using twitter; however, its flexible architecture makes it possible to apply it to other Spanish-speaking countries and allows its integration with other social networks.
**Keywords:** Segmentation, profiling, social network, natural language processing

## *1 Introducción*

Conocer de forma ágil las características, necesidades y preferencias de los consumidores se ha convertido en una labor que requiere cada vez más agilidad ya que los veloces cambios en los mercados obligan a las organizaciones a ser cada vez más flexibles y a adaptarse a las necesidades de los consumidores en tiempos cortos. Considerando lo anterior, diferentes empresas e investigadores han centrado su atención en conocer las características de los individuos a partir de sus comportamientos en redes sociales (Rangel et al., 2015). Si bien ya se han logrado importantes avances en esta labor en idiomas como el inglés, los avances en otros idiomas aún son incipientes (Rangel, 2015).

Con este fin, y en el marco de la alianza CAOBA (Alianza CAOBA, 2017) se desarrolló el Sistema de Segmentación y Perfilamiento Digital (SSPD) que, mediante el uso de técnicas de Procesamiento de Lenguaje Natural (PLN), análisis de redes de grafos y técnicas de aprendizaje automático, permite generar para

ISSN 1135-5948 

cada uno de los usuarios de la red social su perfil e identificar el segmento al que pertenece. Todos los resultados de este sistema se pueden observar mediante un modelo de visualización que plasma el comportamiento y las características de los usuarios digitales, así como también de los segmentos en un periodo determinado o su evolución en el tiempo. Este proyecto es una aplicación industrial de PLN que hace uso de desarrollos y herramientas lingüísticas.

## *2 Desarrollo del sistema*

El desarrollo del proyecto se realizó mediante una arquitectura tipo *SOA* (*Service Oriented Architecture*), donde diferentes componentes se comunican mediante la transferencia de datos en un formato debidamente definido o mediante la coordinación de dos o más servicios (Bell, 2009). En la figura 1 se puede observar el flujo de información asociado al proyecto.

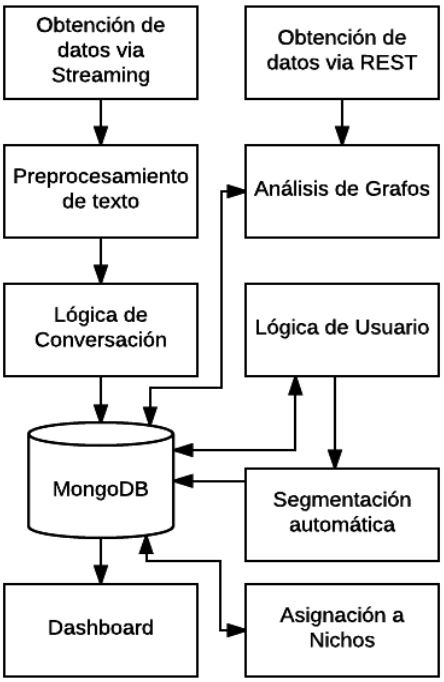


Figura 1: Flujo de información

### 2.1 Obtención de Datos via Stream y Rest

La primera etapa en el proceso del desarrollo del sistema consistió en obtener y almacenar la información de los usuarios digitales. Para la primera versión se seleccionó twitter como la fuente de información sobre los usuarios, aunque para el desarrollo de los componentes se tomó en cuenta la versatilidad necesaria para incluir otras redes sociales.

Para este proceso se hizo uso de dos de las APIs suministradas por *Twitter* (API overview, 2017). Una de ellas, *Streaming*, se usó para obtener en tiempo real las conversaciones (tuits) que tenían su origen en Colombia. Esta *API* también permitió el acceso a información asociada al perfil del individuo.

La segunda API usada fue de tipo *Rest* y se usó para obtener las *timelines* de usuarios que se consideran influyentes en la red; esta información se usó para la elaboración de grafos. La información proveniente de estas dos APIs se almacenó en una base de datos MongoDb.

### 2.2 Preprocesamiento de texto

Posterior al proceso de obtención de la información, tanto los textos de las conversaciones, como las descripciones de los usuarios, pasaron por un proceso de preprocesamiento. En esta etapa se realizó un proceso de tokenizado, *stemming* y *Part of Speech Tagging* haciendo uso de la librería NLTK (Natural Language Toolkit, 2017).

### 2.3 Derivación de Variables

Posteriormente, se realizó la caracterización de los usuarios. Este proceso se dividió en dos lógicas, una enfocada en la conversación, y otra enfocada en el usuario. En la figura 2 se observa un subconjunto de las variables inferidas para el perfilamiento del usuario.

La lógica de conversación se ejecuta en tiempo real, mientras que los procesos subsiguientes se ejecutan al finalizar cada mes, y procesan la información recolectada durante el periodo.

#### 2.3.1 Lógica de conversación

Durante la lógica de conversación se procesaron los datos obtenidos mediante la *API* de *streaming*. En esta etapa el identificador usado fue el ID de la publicación.

En este proceso se realizaron la mayoría de actividades relacionadas con el PLN, particularmente en las variables de Polaridad, Tema, Emoción y Habla Sector:

- La variable Polaridad identifica si la conversación tiene una carga positiva, negativa o neutral. Este proceso hace uso de una metodología *Bag of Words* con lematización,

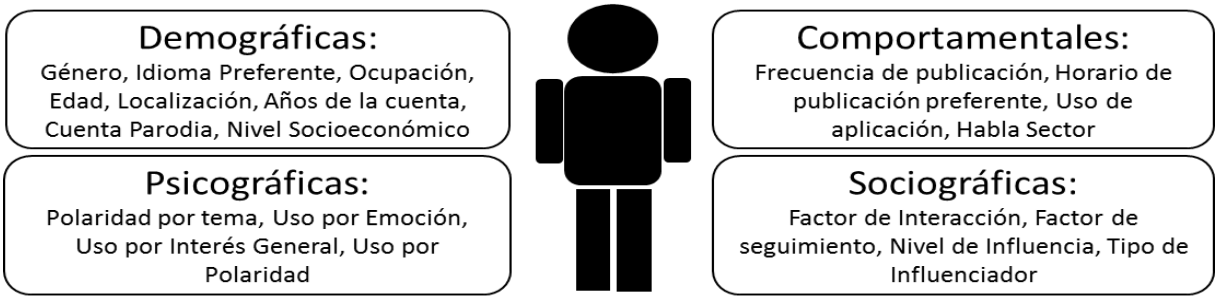


Figura 2: Perfilamiento del Usuario

que calcula la polaridad de la conversación a partir de un lexicón desarrollado en el proyecto. Este lexicón fue producto de un ensamble entre seis lexicones de uso general disponibles en español (Moreno et al., 2017).

- La variable Tema identifica si en la conversación se está hablando de algún tema previamente definido. Para este proceso se realizaron listas de palabras (y lemas) fuertemente asociadas a los temas, y se generó un conjunto de entrenamiento basado en un alto volumen de *hashtags*. Dependiendo de los *hashtags* y de las palabras usadas en la conversación, se asignó el tema.
- La variable Emoción, de una manera similar a lo realizado en la variable polaridad, también hace uso de un enfoque basado en *Bag of Words* con lematización. El lexicón usado fue el *Spanish Emotion Lexicon* (Rangel et al., 2014), el cual asigna a diferentes palabras una emoción y una fuerza de pertenencia.

Adicionalmente, para esta labor se construyó un lexicón donde se asignaba a cada una de las emociones diferentes emoticones y hashtags. De tal manera que, para el proceso de asignación de la emoción, se hizo uso de una métrica que tomaba en cuenta estos diferentes tipos de atributos.

- La variable "Habla Sector" tiene la función de detectar si la conversación en cuestión hace referencia a algún tema diferente a los incluidos en la variable tema, pero relevante para una industria en particular.

Para esto hace uso de una taxonomía en la cual se tienen divisiones por el tipo de relación que tiene la palabra con el tema. Haciendo uso de esta taxonomía, se registra si el tuit habla del sector, y en caso afirmativo, registra de qué manera (Lugar, ocasión, etc.).

### 2.3.2 Lógica de usuario

Por otro lado, la lógica de usuario tuvo dos papeles fundamentales: se encargó de generar un agregado por usuario de los resultados de la lógica de conversación, y realizó aquellos análisis en los cuales se tomaba información directamente relacionada al usuario.

En esta etapa, dos de las variables hicieron uso de PLN para detectar o inferir características del usuario: Género y Nivel Socioeconómico.

- La variable género hizo uso de varios recursos para inferir el género del dueño de la cuenta. En un principio se hace uso de una lista de nombres con su género respectivo, el cual se cruza con el nombre asociado a la cuenta, o en su defecto con el identificador de la cuenta. Si se encuentra más de un nombre, se asigna el género del nombre con más caracteres. En caso de que el nombre no sea encontrado, se procede a analizar la morfología de las palabras presentes en la descripción de los usuarios, y a partir de ella asigna el género correspondiente.
- La variable Nivel Sociocultural, categoriza al individuo en una categoría (alta, media o baja) a partir de las profesiones o cargos que encuentra en la descripción del usuario. Para esto se hace uso de las profesiones lematizadas, las cuales fueron previamente categorizadas según los ingresos esperados.

## 2.4 Lógica de Segmentación

La lógica de segmentación buscó agrupar a los usuarios en diferentes subgrupos según sus características. Para ello, se tomaron varias aproximaciones, incluyendo Análisis de Grafos, Asignación a Nichos Predefinidos y Análisis de Segmentos Automáticos.

El proceso de análisis de grafos se usó para identificar comunidades de temas a partir de coocurrencias de *hashtags* mediante diagramas de Voronoi (Okebe, Boots, y Sugihara, 1992). Posteriormente, mediante un análisis de temática realizado a estos hashtags mediante Alchemy, una API de Bluemix (IBM, 2017), se procedió a caracterizar cada una de las comunidades.

Para la asignación de nichos predefinidos se asignó al usuario un posible nicho de mercado predefinido por una organización. Para realizar esta asignación se generó una lista de palabras (y lemas) que usaría una persona que pertenezca al nicho y a su vez, se les asignó una fuerza de pertenencia. Adicionalmente, para el cálculo se toma en cuenta la polaridad de la conversación, de tal manera que, si la persona está hablando negativamente de las palabras del nicho, en vez de acercarse, se alejará.

Posteriormente, para el proceso de Análisis de Segmentos Automáticos se tomaron los usuarios asignados a cada uno de los nichos, y con cada uno estos grupos se realizó un proceso de *clustering*. El objetivo de este proceso fue facilitar la identificación de subgrupos de usuarios en cada nicho, con lo que se mejoraría el entendimiento de los usuarios asociados.

Para el proceso de *clustering* se incluyeron variables de tipo demográfico y psicográfico. El algoritmo empleado para el proceso fue *Self Organizing Maps (SOM)*, usando el paquete Kohonen de R (Wehrens, 2015).

## 2.5 Dashboard

Finalmente, los resultados de las diferentes lógicas fueron plasmados en un tablero dinámico. Este tablero permite ver de manera gráfica los descriptivos y realizar algunos tipos de consulta. El enfoque seguido para la formulación del tablero se basó en las necesidades particulares del negocio.

Para la construcción de este tablero se hizo uso de Angular 2, junto a Node.js, TypeScript, JavaScript, JQuery y HTML 5. Por otro lado, para la construcción de los gráficos se hizo uso de Highcharts y Echarts.

## *3 Conclusiones*

Este artículo describe un sistema para el perfilamiento y segmentación de usuarios digitales. El sistema fue diseñado de manera flexible para que pueda adaptarse con facilidad a otros países de habla hispana, a otros sectores empresariales y a múltiples redes sociales abiertas.

En este proyecto se hace uso de múltiples recursos de PLN y minería de datos que permiten conocer a los usuarios desde diferentes perspectivas, lo cual es de gran valor para una organización ya que le permite tomar decisiones informadas.

## *Reconocimientos*

Este proyecto fue ejecutado por el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA). El cual se encuentra liderado por la Pontificia Universidad Javeriana (Colombia) y financiado por el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia (MinTIC).

## *Bibliografía*

Alianza CAOBA, 2017. Disponible en http://alianzacaoba.co/. Recuperado 7 de marzo 2017.

API Overview, 2017. Disponible en https://dev.twitter.com/overview/api Recuperado 7 de marzo 2017.

Bell, M., 2009. SOA modeling patterns for service oriented discovery and analysis. John Wiley & Sons.

IBM. 2017. *Alchemy Language*. Disponible en https://www.ibm.com/watson/developercloud/alchemy-language.html. Recuperado 7 de marzo 2017.

Moreno, L., P. Beltrán, J. Vargas, C. Sánchez, A. Pomares, J. Alvarado y J. García. 2017. CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis. En *Proceedings of the 19th International Conference on Enterprise Information Systems -Volume 1: ICEIS*, páginas 288-295.

Natural Language Toolkit, 2017. Disponible en www.nltk.org Recuperado 7 de marzo 2017.

Okebe, A., B. Boots y K. Sugihara, 1992. Concepts and applications of Voronoi diagrams. Il Wiley. New York.

Rangel, F., P. Rosso, M. Potthast, B. Stein y W. Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015, En *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, páginas 1-8.

Rangel, I., S. Guerra y G. Sidorov. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, *29*(1):31-46.

Wehrens, R., 2015. Package 'kohonen'. Disponible en https://cran.r-project.org/web/packages/kohonen/kohonen.pdf.